# George: A visual analytics platform for social media applications

A Thesis
Presented to
The Academic Faculty

by

Damilola F. Animashaun

In Partial Fulfillment
of the Requirements for the Degree
Bachelors of Science in the
College of Computer Science

Georgia Institute of Technology
May 2017

George: A visual analytics platform for anonymous social media applications

Approved by: Dr. Amy Bruckman
Advisor College of Computing
Georgia Institute of Technology

Approved by: Dr. Alex Endert
Advisor College of Computing
Georgia Institute of Technology

Date Approved: May 5th, 2017

# TABLE OF CONTENTS

# INTRODUCTION

Homophily is the tendency of individuals to associate and bond with others who are similar. There is a growing body of work that examines the effect of this phenomenon on online communities. Research has examined the network structure of users in online communities and its effects on discourse and engagement [1, 9]. However, information visualization tools that provide a community-focused approach to content generated by these communities are relatively scarce.

Textual information visualization tools and systems that are geared toward social media applications in general are common, however, the vast majority of these tools tend to display the data from an aggregate instead of a community-focused perspective [2, 5]. There are a few tools that leverage geo tagged content to provide a much more faceted view of the data [3, 4]. They are a step in the right direction but the vast majority of social media data is not geotagged thus they are only applicable to a small subset of social media applications.

Location based social media application Yik Yak launched in 2013. It allows users to post anonymously to a thread that is only visible to others who are in the user's vicinity. It also allows users to upvote and downvote posts by others. It is extremely popular on college campuses. The anonymity provided by Yik Yak often results in candid discussion that would not occur if the user's identity were associated with the conversation. The anonymity also makes it easier to engage in abuse and bigotry. Due to the fact that posts can only be created and voted upon by users in the same vicinity the

content of Yik Yak is somewhat representative of the views and interests of users within a geographic locale.

George, named after Georgia Tech's fictitious George P Burdell, began as a tool to analyze Yik Yak feeds across different colleges. The types of questions that George is intended to help users answer are:

- What topics have people been discussing recently?
- What is the overall sentiment towards a topic?
- How is discussion about a specific topic similar or different across communities?
- How has discussion of a topic evolved over time?

After a few iterations we realised that these questions are also applicable to other forms of online communities. George uses keyword extraction and sentiment analysis techniques to power a data visualization dashboard for exploring textual content. It seeks to empower users to tease out the differences and similarities between different kinds of online communities. By providing a schematic interface, George is database and medium agnostic. As long as the data conforms to George's expectations it will visualize textual data from a variety of data sources. The objective of George is to serve as an initial foray into and spark discussion about the current state of corpus visualization tools for online communities.

# METHODS

## *Data Pipeline*

The core idea behind George is the idea of a unified schema for messages regardless of their origin. A message in George is composed of a timestamp, a community label and a score. The score is computed differently for each datasource. For the prototype our primary data sources were Twitter, Reddit, and Yik Yak. For messages obtained from Twitter, the score is simply the number of retweets. On Reddit and Yik Yak it is the difference between upvotes and downvotes that the message received.

Using the service's respective API's we periodically pulled messages from each data source. At the time of ingestion we also perform sentiment analysis on the individual messages. We use VADER, a sentiment analysis tool that is geared towards the style of writing often seen in social media [12]. It also works for texts from other domains. Using VADER we extract a positive, negative, and neutral score for each message. Other sentiment analysis methods may be used with George provided they output the three pieces of metadata. The messages are then saved with this additional metadata. We use Elasticsearch as our datastore. It is the leading information retrieval database providing advanced features for full text search and aggregation.

## *Keyword Extraction*

There are several heuristic based approaches for extracting keywords from documents. The most popular of these are TFIDF and its variants [6]. TFIDF is the

product of term frequency and the inverse document frequency. The frequency of a term is the number of times that the document occurs in the document. The inverse document frequency is the logarithmically scaled inverse fraction of the documents that contain the word. TFIDF is not suitable for the predominantly short text of social media due to its sparse and noisy nature.

Instead we use TextRank [7], a graph based ranking model for text processing. Graph based ranking algorithms generally seek to determine the importance of a vertex in a graph. Our "graph" in this case is a semantic network derived from our messages. Each term is represented by a node. Each node is connected to other nodes whose term occurs within 10 tokens of itself. In order to generate the keyword graph we first query for all the messages that were generated within the same time period. Then we construct a semantic network from the messages. We run the TextRank algorithm on the network in order to determine the ranking of each node and select the top 10 as the most relevant keywords. We then return this graph to the client and it is rendered as a flat and linearized network in which the nodes are ordered by the score assigned by TextRank.

### *Application Architecture*

George adopts a traditional 3-tier architecture. The presentation tier is a single page web application that provides two different views. The logic tier is a Python web server that exposes a RESTful (Representational State Transfer) API for the presentation tier to communicate with. It is at this level that computationally intensive tasks like keyword extraction and aggregations are performed. The server is partitioned into several different services that provide data for a specific visualization or interaction on the client.

One of the main goals of George is to be a modular platform that can be extended to work with varying data sources and platforms. In keeping with this goal we adopt the Task Pattern for handling requests sent from the client. The Task Pattern is an object oriented design technique that encapsulates a well defined action. With regards to George each request handler invokes a series of Tasks that it needs in order to get the data for each request. The primary advantage of this is that it decouples the incoming request from the specific steps needed to create a response. For example one can now use a different database or adopt a different keyword extraction technique without having to make any changes to the rest of the application.

# RESULTS

## Dashboard View

The Dashboard View serves as the main entry point to the application for a collection of communities. It is intended to help users answer the following questions:

- What topics have people been discussing recently?
- What is the overall sentiment towards a topic?
- How is discussion about a specific topic similar or different across communities?

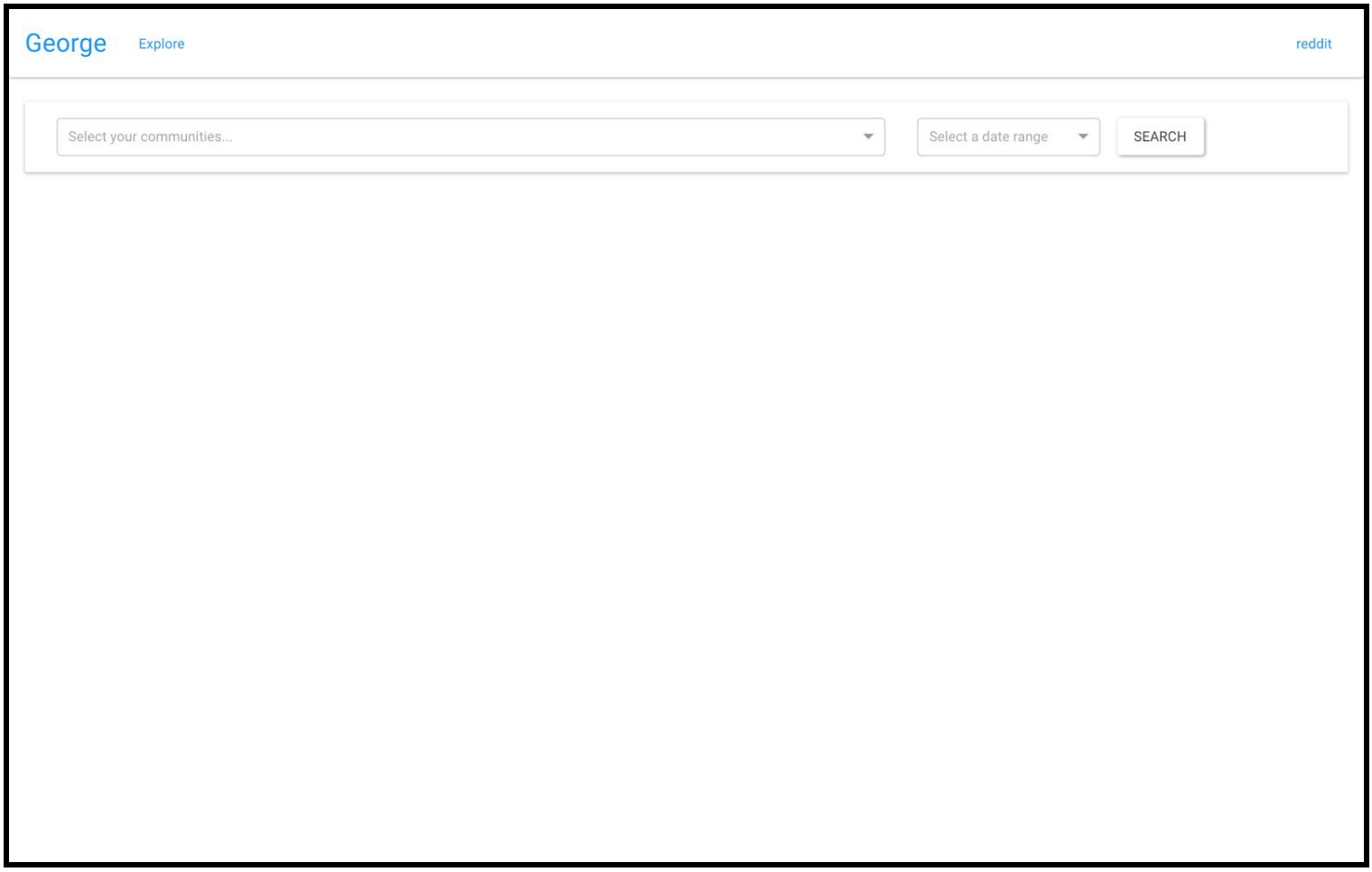


*Figure 1*

Figure 1 shows the initially empty view that has yet to be populated with data. The user has a multi select drop down to pick the communities that they are interested in and a drop down to to specify a time period(past 4 hours, past day, or past 7 days).

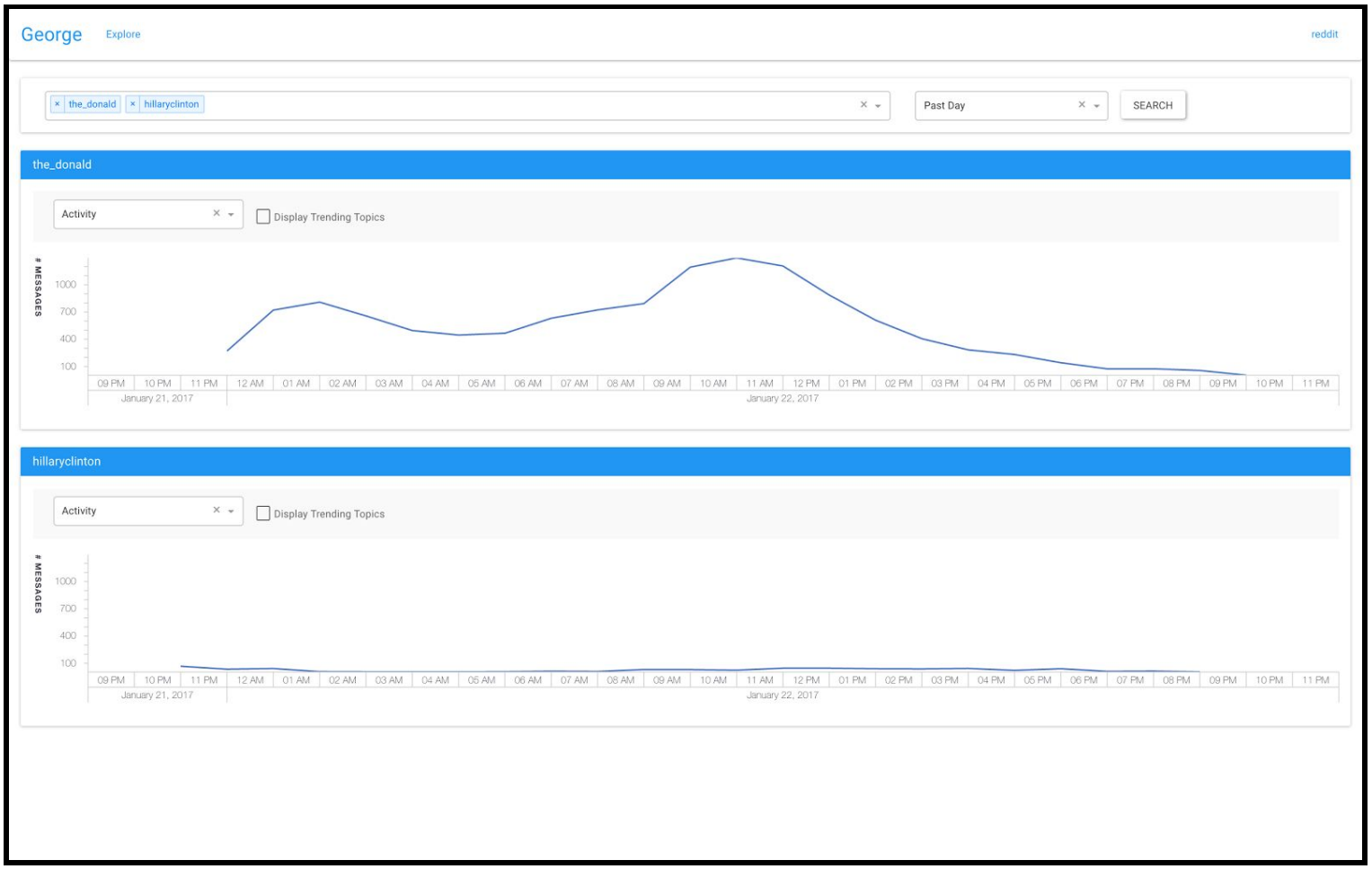


*Figure 2: Dashboard View displaying activity data*

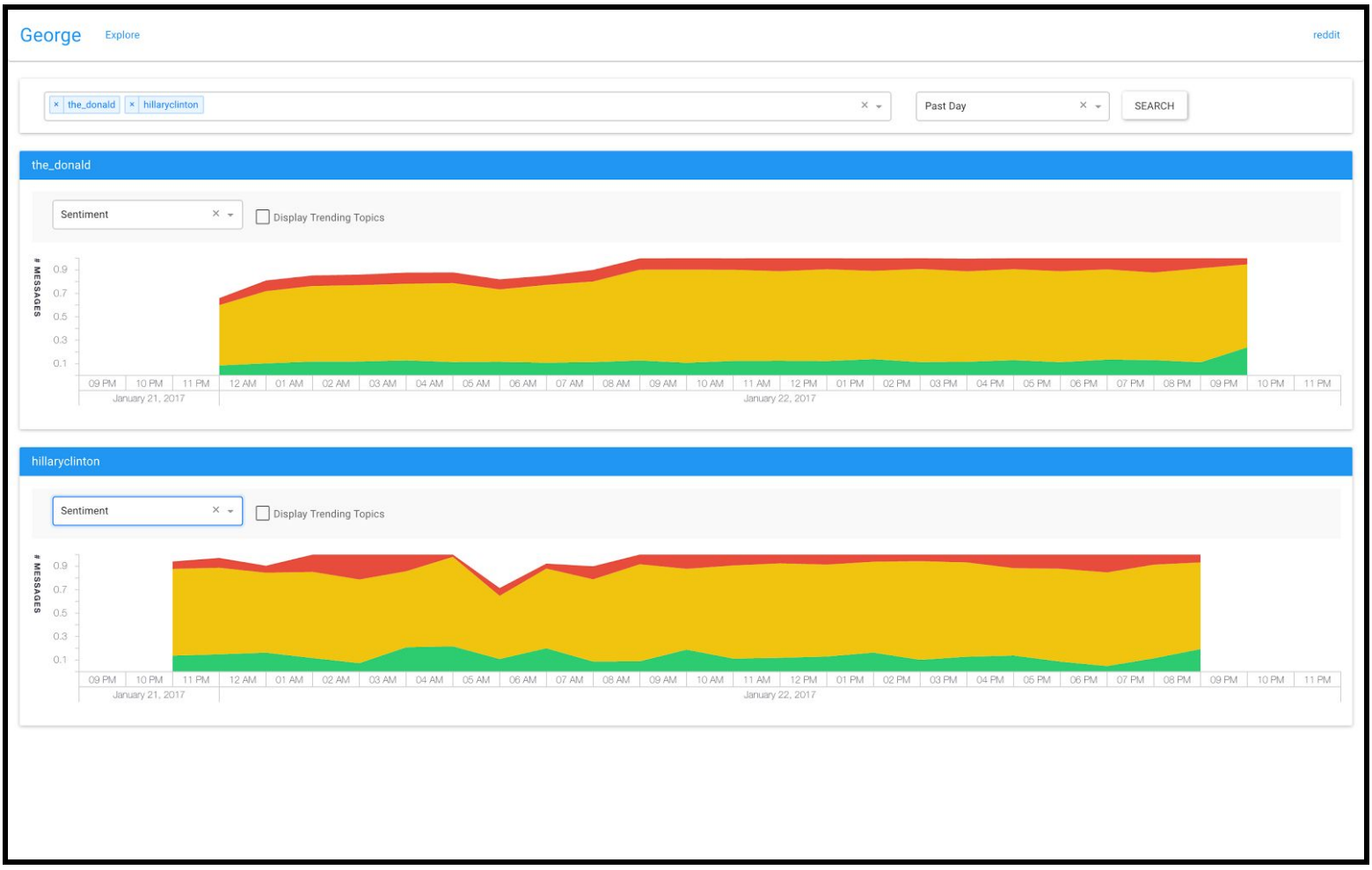


*Figure 3: Dashboard view displaying sentiment data.*

Once the user hits search the screen transitions to a list view. Each selected community has a panel that displays line graph that charts timeline of selectable attributes for the community. The choices are activity, the voting pattern, and sentiment. The graph for activity and voting pattern are simple line charts with time on the x axis and the corresponding attribute on the y axis(Figure 2) . The sentiment chart is a stacked area chart with 3 distinct areas: green for positive, yellow for neutral and red for negative(Figure 3).


*Figure 4: Dashboard view with trending topics enabled.*

The user can click a checkbox to display the trending topics associated with each community. By default the data returned encompasses the currently selected time interval. The user can specify a time interval by clicking on the beginning of their desired interval, dragging to the end and releasing the mouse. The trending topics graph changes to display topics from the interval. When the user clicks on a node a tooltip appears that shows instances of the keyword in context.

## Explore View


The Explorer View can be accessed by clicking on the explore link located in the navigation bar. This view is intended to help users determine how usage specific terms and phrases differ from community to community in terms of interest, voting activity, and sentiment. The user can select up to four terms or phrases, communities that they wish to search, and the time period that they wish to filter by .Each search results in the display of a panel for each search term. Each chart is composed of two graphs. The first is a multi line chart that shows the activity of the term over the specified period of time. The second chart is a bar chart that shows the average frequency or sentiment over the specified time interval.

## Scenario

In this section we will demonstrate how George can be used to quickly and efficiently explore similarities and differences among several online communities. For this discussion our fictional user will be a social media intern for a statewide political campaign. Our user is tasked with keeping an eye on several politically focused online forums. Suppose a negative story has recently surfaced about the candidate that our user works for. To learn how the story is being perceived by supporters and opponents alike our user starts up George and navigates to the Dashboard View. On the Dashboard View the user selects several forums that are of particular interest and sets the filter to only include posts from the past day. As expected the number of posts on each forum has increased sharply since the story broke. The user is particularly interested in how the story is being perceived on a forum that is primarily composed of people who are undecided about which candidate to vote for. In order to determine the topics that are being discussed the user enables trending topics and selects the interval that includes the recent spike in activity. The trending topics appears and displays topics from the selected interval. Most of the displayed topics are about details from the negative story. The size of each node indicates that some topics about the story are being discussed much more frequently than others. Variations in the thickness of the edges between several nodes indicate which topics are being discussed in tandem. The user repeats the process for forums populated primarily by supporters and opponents. By comparing the individual trending graphs for each forum our user is able to determine which aspects of the story are of particularly prominent for each forum and by extension group of voters.

## Usage Observation

In order to evaluate and improve the design of George we performed an informal observation of a small group of users. We were particularly interested in gauging George's usability and effectiveness in helping users discern differences between the content of different online forums. The participants were 3 females and 1 male between the ages of 20 and 22. They used George to explore the subreddits of the presidential candidates of the 2016 election(r/The_Donald, r/HillaryClinton, r/JillStein, r/GaryJohnson). In order to organically tests usability users were not provided with tutorials or given specific tasks. They were simply asked to explore the application and to think aloud as they played with it. After 5 minutes of interaction users were given a quick tour of all the available features and were asked to discuss their opinions on the application.

Each trial began in a somewhat similar manner. Users would begin at the Dashboard View where they would select the communities and time period that they were interested in. Once the cards loaded the users would examine the cards in sequence. Afterwards they would return to the first card and start to interact with the drop down to toggle the data that was displayed on the line graph. Users easily grasped the mapping of colors to sentiments. Votes were a bit problematic. Some users were not familiar with how Reddit's voting mechanism worked so they needed additional clarification as to what the view depicted. Since the idea of votes may differ slightly from community to community(for example if Twitter were the data source then retweets could be mapped to votes) this particular selection may require additional clarification in the user interface

depending on the data source. After taking in the line graphs displayed on the individual cards users would often select the option to display trending topics. In order for the graph of topics to display users would have to select an interval on the line graph. Since this was unknown to the users they would just wait for additional data to appear on screen instead of selecting an interval. Once users managed to get the trending graphic to display they found the resulting graph relatively intuitive to reason about(larger nodes correspond to more frequent topics and thicker edges correspond to greater co-occurrence).

None of the users navigated to the Explore View without additional prompting. The link to it is easy to miss since it is located in the navigation bar. Since the input fields are identical to those in the Dashboard View the users attempted to use it in the same manner. That is they neglected to enter search terms to query. This occurred frequently despite the error message displayed whenever they attempted to launch the view. Once prompted users were quickly able to determine how their input mapped to the data that was being displayed.

As a result of these informal observations we found that users could easily understand the data being displayed but experienced difficulty in performing the correct sequence of actions to bring up the data. Features like the trending topics graph and the input to the Explore View were not discoverable through experimentation but were greatly appreciated once explained. A potential solution to this issue is to add tooltips containing helpful information to each input and to provide a brief sequenced tutorial for each view. Ideally the tutorial should take the user through every interaction method in each view and how it affects the data that will be displayed. This is similar to the

onboarding process of most popular consumer web applications. In conclusion, users praised the simplicity of George but requested additional help in learning how to use it.

## DISCUSSION AND FUTURE WORK

George is an initial foray into creating corpus visualization tools geared towards online communities. It provides a few basic features for investigating the content of online communities. The logical next step would be to perform a formal user study to determine the usability of the prototype, its ability to foster discovery and exploration, and any additional features that may prove useful.

One aspect that George does not consider are the individuals who form online communities. Node-link diagrams are the de facto method for visualization network structure in online communities. Systems like Vizster and SocialAction pair node-link diagrams with coordinated views in order to support visual search and analysis of online social network structure [10, 11]. While these systems do a superb job of highlighting network structure, they are not suitable for other germaine tasks. They do not facilitate connecting individuals with the content that they generate. Nor can they compare an individual's content to the overall content of the community that they are part of and its reception by the community. This would be particularly relevant for identifying thought leaders and frequent dissenters. Thus a relevant extension to George would be to add features that allow users to explore the contributions of individual members within a community.

# REFERENCES


1.) Yardi, Sarita, and Danah Boyd. "Dynamic debates: An analysis of group polarization over time on twitter." Bulletin of Science, Technology & Society 30, no. 5 (2010): 316-327.

2.) Hu, Mengdie, Krist Wongsuphasawat, and John Stasko. "Visualizing Social Media Content with SentenTree." IEEE Transactions on Visualization and Computer Graphics (2016).

3.) Yin, Jie, Andrew Lampert, Mark Cameron, Bella Robinson, and Robert Power. "Using social media to enhance emergency situation awareness."*IEEE Intelligent Systems* 27, no. 6 (2012): 52-59.

4.) Xia, Chaolun, Raz Schwartz, Ke Xie, Adam Krebs, Andrew Langdon, Jeremy Ting, and Mor Naaman. "CityBeat: real-time social media visualization of hyper-local city data." In Proceedings of the 23rd International Conference on World Wide Web, pp. 167-170. ACM, 2014.

5.) Wu, Yingcai, Nan Cao, David Gotz, Yap-Peng Tan, and D. Keim. "A survey on visual analytics of social media data." IEEE Trans. Multimedia (2016).

6.) Lott, Brian. "Survey of Keyword Extraction Techniques." UNM Education (2012).

7.) Mihalcea, Rada, and Paul Tarau. "TextRank: Bringing order into texts." Association for Computational Linguistics, 2004.

8.) Perrin, Andrew, M. Duggan, L. Rainie, A. Smith, S. Greenwood, M. Porteus, and D. Page. "Social media usage: 2005-2015. Pew Research Center." (2015).

9.) Bisgin, Halil, Nitin Agarwal, and Xiaowei Xu. "Investigating homophily in online social networks." In Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on, vol. 1, pp. 533-536. IEEE, 2010.

10.) Heer, Jeffrey, and Danah Boyd. "Vizster: Visualizing online social networks." In IEEE Symposium on Information Visualization, 2005. INFOVIS 2005., pp. 32-39. IEEE, 2005.

11.) Perer, Adam, and Ben Shneiderman. "Balancing systematic and flexible exploration of social networks." IEEE Transactions on Visualization and Computer Graphics 12, no. 5 (2006): 693-700.

12.) Hutto, Clayton J., and Eric Gilbert. "Vader: A parsimonious rule-based model for sentiment analysis of social media text." In Eighth International AAAI Conference on Weblogs and Social Media. 2014.